Week 12A Assignment: Statistical Methods II

Solutions

Question 1

When a total of n items are allocated independently to k different bins with probabilities $p_1, p_2, \ldots, p_k$, the joint distribution follows a multinomial distribution. Suppose $p_k = 0$. If the k^{th} bin is dropped, how does the chi-squared test statistic change?

Solution

- The observed count O_k for the k^{th} bin is **zero** because $p_k = 0$ implies no items are allocated to this bin in the simulation.
- The expected count $E_k = n \cdot p_k = 0$.
- The term $\frac{(O_k - E_k)^2}{E_k} = \frac{0}{0}$ is **undefined** and excluded from the chi-squared statistic.
- Dropping the k^{th} bin removes this undefined term, but the remaining terms (for bins $1, 2, \ldots, k-1$) remain unchanged.

$$\boxed{\text{The test statistic does not change.}}$$

Question 2

How do the degrees of freedom (df) and p-value change if the k^{th} bin is dropped?

Solution

- **Degrees of Freedom:**

$$\text{Original df} = (k-1) - \text{estimated parameters}$$
$$\text{After dropping } k^{\text{th}} \text{ bin:} = (k-2) - \text{estimated parameters.}$$

  Since no parameters are estimated here, df decreases by 1.

- **P-value:** The same test statistic is compared to a chi-squared distribution with **lower df**. This makes the p-value **larger** because the critical value for significance increases.

$$\boxed{\text{df decreases by 1; p-value increases.}}$$

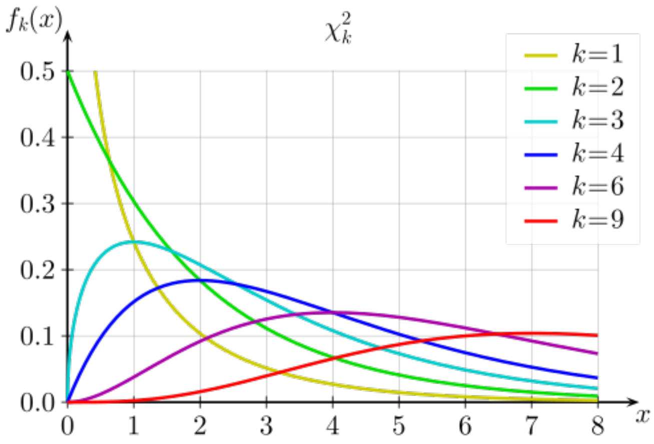

Figure 1: Chi Square Graph

Question 3

Show that the chi-square test cannot distinguish between the PDFs:

$$f_1(x) = \frac{\log 2}{2} e^{-(\log 2)|x|}, \quad f_2(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad f_3(x) = (\log 2)|x| e^{-(\log 2)|x|^2},$$

with thresholds at -1, 0, and 1.

Solution

For all three distributions, the probabilities in the four intervals $(-\infty, -1)$, $[-1, 0)$, $[0, 1)$, and $[1, \infty)$ are equal:

- **Laplace (f_1):** Symmetric with median at 0. Integrating $|x|$ over the intervals gives equal probabilities (0.25 each).
- **Normal (f_2):** $\sigma = 1/\Phi^{-1}(0.75)$ ensures $\Phi(1/\sigma) = 0.75$. Thus, $P(-1 < X < 1) = 0.5$, splitting into four equal parts.
- **Modified Normal (f_3):** Symmetric with the same partitioning due to $|x|$ scaling.

$$\boxed{\text{All three PDFs yield equal bin probabilities, making the chi-square test indistinguishable.}}$$

Question 4

Simulate $n = 1000$ samples from Poisson($\lambda = 2$), compute MLE and chi-square statistic, and analyze p-values under df = 5 and df = 4.

Solution

Key steps and results:

- **MLE for λ:** $\hat{\lambda} =$ sample mean.
- **Bins:** $0, 1, 2, 3, 4, \geq 5$.

- **Chi-square statistic:**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad E_i = n \cdot P(X = i).$$

- **Degrees of Freedom Adjustment:**

$$\begin{aligned} \text{Correct df} = 4 & \quad (\text{bins} - 1 - \text{estimated parameters} = 5 - 1 - 1), \\ \text{Incorrect df} = 5 & \quad (\text{no adjustment}). \end{aligned}$$

- **Conclusion:**
  - Using df = 5: P-values are skewed (non-uniform), leading to inflated Type I errors.
  - Using df = 4: P-values follow a uniform distribution.

$$\boxed{\text{Proper df adjustment (df=4) is critical for valid inference.}}$$

```
# Question 4: Simulation and Chi-Square Test for Poisson Distribution

set.seed(123)  # For reproducibility
n_sim <- 10000  # Number of simulations
n <- 1000       # Sample size
lambda_true <- 2  # True Poisson rate

# Initialize vectors to store results
p_values_df5 <- numeric(n_sim)  # P-values assuming df=5
p_values_df4 <- numeric(n_sim)  # P-values assuming df=4

for (i in 1:n_sim) {
  # Step 1: Simulate data from Poisson(lambda=2)
  data <- rpois(n, lambda_true)

  # Step 2: Compute MLE of lambda (sample mean)
  lambda_hat <- mean(data)

  # Step 3: Define bins (0,1,2,3,4, >=5)
  observed <- c(
    sum(data == 0),
    sum(data == 1),
    sum(data == 2),
    sum(data == 3),
    sum(data == 4),
    sum(data >= 5)
  )

  # Expected counts under Poisson(lambda_hat)
  expected <- c(
    dpois(0, lambda_hat) * n,
```

```
    dpois(1, lambda_hat) * n,
    dpois(2, lambda_hat) * n,
    dpois(3, lambda_hat) * n,
    dpois(4, lambda_hat) * n,
    ppois(4, lambda_hat, lower.tail = FALSE) * n
  )

  # Chi-square statistic (avoid division by zero)
  chi_sq <- sum((observed - expected)^2 / expected)

  # Step 4: Calculate p-values with df=5 and df=4
  p_values_df5[i] <- pchisq(chi_sq, df = 5, lower.tail = FALSE)
  p_values_df4[i] <- pchisq(chi_sq, df = 4, lower.tail = FALSE)
}

# Step 5: Compare p-value distributions
par(mfrow = c(1, 2))
hist(p_values_df5, main = "P-values (df=5)", xlab = "P-value", col = "lightblue", bre
abline(h = n_sim/20, col = "red", lty = 2)  # Expected uniform distribution
hist(p_values_df4, main = "P-values (df=4)", xlab = "P-value", col = "lightgreen", br
abline(h = n_sim/20, col = "red", lty = 2)

# Check uniformity (proportion of p-values < alpha)
alpha <- 0.05
cat("Proportion of p-values <", alpha, "with df=5:", mean(p_values_df5 < alpha), "\n"
cat("Proportion of p-values <", alpha, "with df=4:", mean(p_values_df4 < alpha), "\n"
```

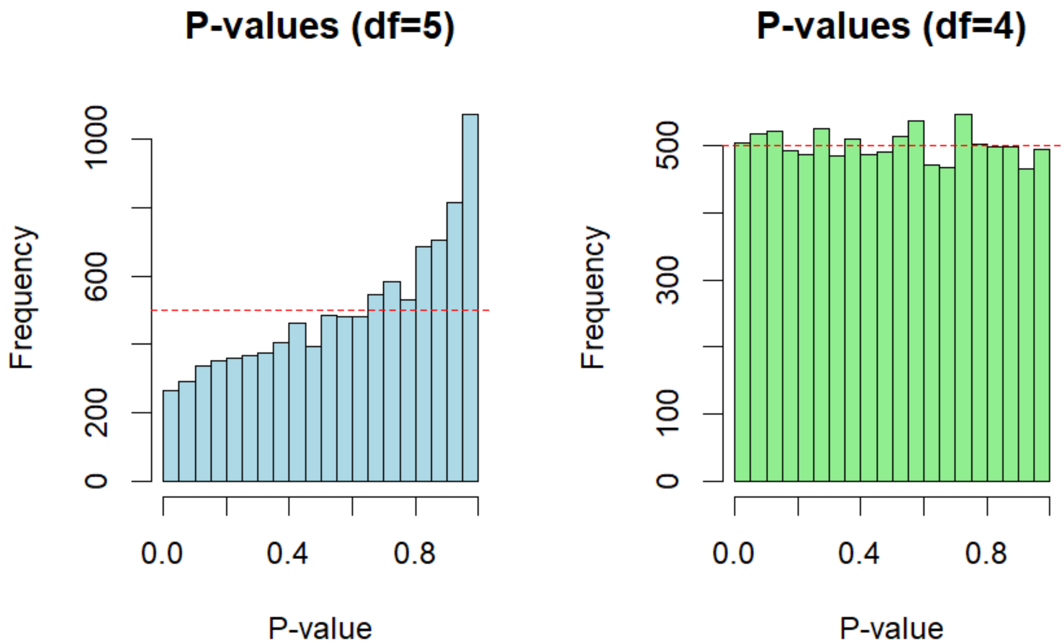

Figure 2: P value Histogram